SWAMP: AN ISOMETRIC FRONTEND FOR SPEAKER CLUSTERING

Patrick Nguyen

Panasonic Speech Technology Laboratory (PSTL)
Panasonic Technologies Company
3888 State Street, Suite 202,
Santa Barbara, CA 93105, U.S.A.
nguyen@research.panasonic.com

ABSTRACT

In this paper, we describe a non-linear feature normalization based on Riemannian differential geometry. This feature normalization will yield parameters that are invariant under any bijective stationary transformation. Moreover, it is robust to additive noise that is uncorrelated with speech and quasi-stationary. The only requirement is that of ergodicity. The frontend is called SWAMP (Sweeping Metric Parameterization).

The frontend assumes that speech resides in a small, smooth manifold that is entirely and densely explored during the course of an utterance. It first observes the tangent spaces on every point of the manifold. This defines a local Riemannian geometry. Under this geometry, we are able to measure geodesic lengths on the manifold. These lengths are invariant under non-linear transformations. Therefore, we are able to locate a point invariantly by measuring its relative distance to all other observed points. Through classical multi-dimensional scaling, we map this triangulation to a canonical, Euclidean, isometric space inherent of the observed manifold.

Combined with standard features, SWAMP features are shown to improve speaker clustering on Broadcast News.

1. INTRODUCTION

State-of-the-art speaker clustering is chiefly based on variations of speech recognition features. It is usually argued that such features are inadequate for the goal of distinguishing between speakers.

In this paper, we consider the audio stream as a solid. The solid's local characteristics are normalized via devices of differential geometry. The algorithm presented herein relies on tried and true mathematical methods of physics of relativity and psychology. In this sound mathematical framework, we aim at achieving two of the most desirable properties among frontends: acquisition channel independence, and noise-robustness.

Intuitively, the algorithm will re-write reference points in a Euclidean space that is naturally deduced from the manifold structure painted by the speech. To that end, it enables devices from Riemannian differential geometry, numerical analysis, graph theory, and optimal dimensionality reduction.

2. THE ALGORITHM

2.1. Preliminary definitions

We start with a few definitions. We observe the signal $x(t)$, or x_t, $t \in [0;T]$, of dimension D. The final reparameterized signal will be $s(t)$ of dimension E. First, define the *time-average* of a function f:

$$\langle f \rangle_p \triangleq \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathrm{d}t \left\{ I_p(x_t) f \right\} \tag{1}$$

where $I_p(\cdot)$ is the indicator function of an infinitesimally small region around p. Unless otherwise stated, we will assume *ergodicity* throughout:

$$\langle f \rangle = \mathrm{E} f, \ \forall f, \tag{2}$$

where $\mathrm{E}(\cdot)$ is the expectation.

The *total difference* theorem is useful in proving properties:

$$\mathrm{d}y = \sum_k \frac{\partial y}{\partial x^k} \mathrm{d}x^k = \frac{\partial y}{\partial x^k} \mathrm{d}x^k, \tag{3}$$

where the second equality is in the Einstein summation notation (e.g. [1]), which we use throughout.

We define three mathematical properties that are desirable for frontends: invariance, ergodicity, and noise-robustness. *Invariance* means that under a transformation $\mathbb{R}^D \to \mathbb{R}^D$:

$$x \mapsto x' = x'(x), \tag{4}$$

properties will remain the same. This is also called parameterization-independence. For instance, time axes are independent. Moreover, we say that a property is intrinsic, or *natural*, if there is one unique way of defining it invariantly by the manifold only. For instance, the dimension of a manifold is intrinsic. However, tangent vectors are not intrinsic but the tangent space is. *Ergodicity* ensures that with an ergodic time sequence, the outcome will also be ergodic. For instance, time reference is not ergodic. Finally, *noise-robustness* implies that some properties may be reasonably well estimated when the signal is corrupted with additive noise.

Finally, we define the Jacobian matrix J:

$$[J_{ik}] = \frac{\mathrm{d}x'^i}{\mathrm{d}x^k}.$$

In general this matrix is non-symmetric, but we assume it invertible $|J| \neq 0$.

2.2. Differential Geometry

Differential geometry is concerned about infinitesimal changes around a point in a manifold. In this mathematical field, manifolds are always C^∞ (smooth) and never abstract. There are two convenient ways of visualizing such a manifold of dimension S: firstly, as a solid in S-dimensional space, or secondly, as a surface embedded in an $(S+1)$ space. One of the most important predicates in differential geometry is that of coordinate invariance. It means that differential geometry manipulates vectors in a way that does not depend on a specific choice of coordinate. It also means that two manifolds are indistinguishable if they only differ by a homeomorphism. The latter point is a limitation to which we will come back later. The speech manifold will be called $\mathcal{M}$.

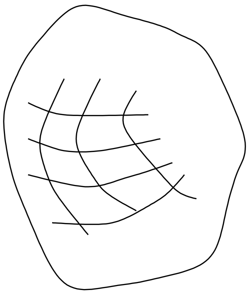

Fig. 1. A manifold with a reference system (grid).

Differential geometry is the device through which we will achieve transformation invariance. Imagine the representation of a manifold as in Figure 1. We will extract entities which are invariant under any deformation of the solid. Deformations correspond to change in coordinates. Coordinates are also non-homogeneous, or local: they depend on the position of the point in the manifold. The space that they span is called the tangent space. The tangent space is a linear approximation of the surface at the point.

Relativistic physics makes heavy use of differential geometry. We introduce two concepts: covariance and contra-variance. They correspond to respectively: going *against* or *in the direction of* variabilities. The original coordinate space is contravariant. Co- and contra-variant are usually represented with lowered and raised indices respectively.

Following work by Levin [2] and Schrödinger [3], we define a Riemannian metric by measuring, at each position p, the directional derivative along the speech trajectory $x(t)$.

We define the *sweeping metric* to be:

$$g^{kj}(p) \triangleq \left\langle \frac{\mathrm{d}x^k}{\mathrm{d}t}\frac{\mathrm{d}x^j}{\mathrm{d}t} \right\rangle_p . \tag{5}$$

This defines a Riemannian contravariant tensor G of type 2: it is a bilinear symmetric, positive definite form with two raised indices. We will omit p where obvious. This tensor can be transformed into a covariant tensor of type 2 by matrix inversion:

$$G^{-1} = [g_{kj}] = [g^{kj}]^{-1}. \tag{6}$$

Length, volume, and angles are invariant of parameterization under this metric. We are primarily concerned with lengths $\mathrm{d}s^2$ of infinitesimal changes $\mathrm{d}x = [\mathrm{d}x^1, \mathrm{d}x^2, ..., \mathrm{d}x^D]^T$:

$$\mathrm{d}s^2 = g_{kj}\mathrm{d}x^k\,\mathrm{d}x^j = \mathrm{d}x^T G^{-1}\mathrm{d}x. \tag{7}$$

It is trivial to see that a change in coordinates will not change under a transformation x':

$$\mathrm{d}s'^2 = dx'^T G'^{-1} dx' = \mathrm{d}s^2, \tag{8}$$

with

$$dx' = Jdx, \quad G' = JGJ^T. \tag{9}$$

Infinitesimal length can be measured invariantly at any point p of the manifold, thanks to the sweeping metric.

We observe that there is a lack of directionality: no vectors can be defined invariantly without further information, only contra- or covariantly. In particular, Principal Component Analysis (PCA) cannot be applied at this stage.

2.3. Triangulation – Geolen reference

The next question to answer is: is there a natural coordinate system associated to the manifold, where we can rewrite the curve $x(t)$ naturally? In this section, we define an invariant coordinate system, which will be extend to a natural coordinate system with multidimensional scaling [4].

We have observed that the time reference is an invariant property of the system. We shall then define a reference system whereby a point is defined by distance with respect to all other points in the manifold: this is a triangulation. Consider two points $p, q \in \mathcal{M}$. Let a curve $\gamma : \{\gamma(t) \in \mathcal{M}\}$, such that $\gamma(0) = p$ and $\gamma(1) = q$. The *geodesic length*, or *geolen* is the shortest line distance [5]:

$$\Delta^2(p,q) = \min_\gamma \int_\gamma \mathrm{d}s^2 \tag{10}$$

Therefore, each point $x(t)$ can be re-parameterized in an invariant reference system:

$$\Delta_k(t) = \Delta[x(t), x(k)], \;\; k = 1, ..., T. \tag{11}$$

Although this measure seems straightforward to define, it is the major practical hurdle.

2.4. Isometric reference system

Although the geolen reference is invariant, it suffers from one major drawback: it is not ergodic. Suppose we repeat the speech twice: the coordinates will be twice as long. A time shift will shift all coordinate indices. This is usually not advisable.

Therefore, the last step of our algorithm enacts a dimensionality reduction technique most popular in psychology. It is called *classical, metric multidimensional scaling*). Given a distance matrix $\Delta^2_{kj} = \Delta^2_k(j)$ from (eq. 11), we can define a doubly-centered distance:

$$B^*_{kj} = -\frac{1}{2}\left[\Delta^2_{kj} - D^{-1}\sum_l [\Delta^2_{kl} + \Delta_{jl}] + D^{-2}\sum_{l,m}\Delta^2_{lm}\right]. \tag{12}$$

The spectral (or eigen) decomposition of B^* is:

$$B^* = U\Lambda U^T. \tag{13}$$

It is truncated as usual to the E largest eigenvectors and eigenvalues. The new coordinate system will define:

$$s(t) = \Lambda_E^{1/2} U_E^T(t). \tag{14}$$

This E-dimensional vector will be the new parameterization. It is natural and ergodic. Because all distances can be computed in a Euclidean homogeneous coordinate system, it is called an *isometric* system:

$$||s(t) - s(\tau)||^2 = \Delta^2(s(t), s(\tau)) = \sum_{k=1}^{E}\left[s_k(t) - s_k(\tau)\right]^2. \tag{15}$$

We come back to the note in the introduction of differential geometry: the coordinate system is defined up to a rotation (homeomorphism). The fundamental axes are defined in the principal directions of energy: they depend on the population of the sampling $x(t)$ of the manifold $\mathcal{M}$. If the sampling is ergodic, then the axes are well-defined. In other words, we are sensitive to the linguistic content of the speech. We hope that silence/speech will help fix the axes. Otherwise, an extrinsic, “oracle” probability measure must resolve the ambiguity.

2.5. Noise robustness

Under certain conditions, the frontend can be robust to additive noise. The noise model is:

$$\tilde{x}(t) = x(t) + w(t). \tag{16}$$

The sweeping metric becomes:

$$\tilde{g}^{kj} = g^{kj} + \left\langle \dot{w}^k \dot{w}^j \right\rangle + \left\langle \dot{x}^k \dot{w}^j + \dot{x}^k \dot{w}^j \right\rangle, \tag{17}$$

where $\dot{f}^k(t) = \frac{\mathrm{d}f^k}{\mathrm{d}t}$ for x and w. A sufficient condition for $\tilde{g}^{kj} = g^{kj}$ is that $\dot{w} \approx 0$, or that the noise be stationary. Similarly, the infinitesimal length becomes:

$$\mathrm{d}\tilde{s}^2 = \mathrm{d}s^2 + \mathrm{d}w^T G^{-1} \mathrm{d}w + 2\mathrm{d}x^T G^{-1} \mathrm{d}w. \tag{18}$$

Noise and speech are assumed uncorrelated so that the third term cancels. Additionally, if the noise contribution is constant on $\mathcal{M}$, then there is a constant bias on Δ which also cancels.

Therefore, quasi-stationary noise which spans a space orthogonal to $\mathcal{M}$ does not corrupt our features.

3. IMPLEMENTATION

The principle of the frontend was shown in the previous section. In trying to extract the quintessence of the algorithm, we have chosen to conceal major practical aspects of the implementation. Most of them are due to the finite nature of the signal. We shall overview them here.

3.1. Quantization

The metric $g^{kj}(p)$ should, in principle, be computed for all points $p \in x(t)$ over an infinite period of time. In practice, it cannot be. Levin proposes to quantize the contravariant space linearly, and then to interpolate tangent spaces. This involves the computation of the derivative of the Riemann metric along a direction m, also called *Christoffel symbol of the second kind*:

$$\Gamma^m_{ij} = \frac{1}{2} g^{km} \left(\frac{\partial g_{ik}}{\partial x^j} + \frac{\partial g_{jk}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^k} \right). \tag{19}$$

Parallel transportation of a tangent vector along a curve γ is:

$$\delta x^j = \mathrm{d}x^j + \Gamma^j_{ik} \mathrm{d}x^i \mathrm{d}x^k. \tag{20}$$

Unfortunately, in general this involves solving large Ordinary Differential Equations. This procedure is numerically unstable. Two problems arise: first, the manifold can be interpolated anywhere on the space; second, the quantization is contravariant and not invariant. The first problem arises if the manifold surface is non-convex, e.g. it is a donut or toroid. Moreover, in [2], the method works well with strongly directive, low-dimensional spaces. Therefore, the curse of dimensionality makes the approach infeasible because the density decreases polynomially with the feature dimension D.

We avoid the computation of tangent spaces at ill-conditioned points altogether by using *vector quantization* to cluster the time points. This is done initially using a contravariant measure of distance, but then it is replaced with $s^2(p,q)$ and quantization iterates. The update of the centroid satisfies:

$$y_c = \arg\min_y \sum_{\mathrm{d}q} \mathrm{d}s^2(y, \mathrm{d}q). \tag{21}$$

It is the empiric mean of the cluster. Our clustering is relatively invariant, that is, under a transformation $x' = f(x)$, if we perform clustering Q', we have:

$$Q'(x' = f(x)) = f(Q(x)). \tag{22}$$

In other words, points are grouped the same in both coordinate systems and f and Q commute. With quantization, our sweeping metric of (eq. 5) defines an invertible metric tensor invertible. It is also possible to have degenerate spaces for certain regions of the manifolds, for instance, where there is silence.

3.2. Geodesics in local metrics

Now we show how to compute the length from any point p to any nearby point q. We assume a piecewise flat structure of the manifold: around each centroid, the doubly covariant tensor g_{kj} is valid. In a very near neighborhood around a centroid c, we have:

$$s^2_c(p,q) = (p-q)^T G^{-1}_c (p-q). \tag{23}$$

Suppose now that we have two regions a and b, we define $V(a)$ the *Dirichlet region* of a to be the set of of points p closest to a:

$$V(a) = \left\{ p : s^2_a(p,a) \leq s^2_b(p,b),\ \forall b \right\}. \tag{24}$$

We call the quadratic hyper-surface for which there is equality the *Voronoi interface* $\pi(a,b)$. The geolen between two points x_a, x_b, one in the $V(a)$, and one in $V(b)$, is a two straight segments intersecting at a point z on the Voronoi interface $\pi(a,b)$. We can write the point z:

$$z = x_a + z_a = x_b + z_b. \tag{25}$$

We can minimize over z:

$$H(z) = s^2_a(x_a, z) + s^2_b(x_b, z), \text{ s.t. } z \in \pi(a,b). \tag{26}$$

We suppose that the are entirely contained in their Dirichlet regions. The other case will be treated later. Using the Lagrangian multiplier λ, we find:

$$z_a = \left\{ (1-\lambda)A + (1+\lambda)B \right\}^{-1} \left[(1+\lambda)B\Delta c + B\Delta x + \lambda(B-A)x_a \right]$$

and in the unconstrained case:

$$z_a|_{\lambda=0} = (A+B)^{-1} B(\Delta c + \Delta x). \tag{27}$$

There is no closed-form solution, but a Newton-Raphson [6] iteration over λ will converge quickly.

3.3. Tunneling

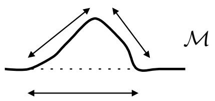

Fig. 2. Tunneling: it is shorter to go "under" the manifold with the dotted line.

As seen on Figure 2, we have to be careful to integrate distances over manifold surface. In (eq. 10), we assume that γ is always on $\mathcal{M}$. Otherwise, we could go "under" a local bump in $\mathcal{M}$ and reducing lengths artificially: this should be avoided. Similarly, a small local dip will introduce an effect called bridging. It is less crucial because of local lengths are then integrated in the graph.

Suppose that we have two points x_1, x_2 in the $V(a)$. The segment is parameterized with $\Phi \in [0;1]$:

$$x(\Phi) = x_1 + \Phi\Delta x = x_1 + \Phi(x_2 - x_1). \tag{28}$$

We perform a *Dirichlet test* to see whether the segment intersects another Voronoi region b:

$$0 < \left[||\Delta x||_a^2 - ||\Delta x||_b^2\right]\Phi^2 + 2\left[\langle \Delta x, x_2 - c_b\rangle_b - \langle \Delta x, x_2 - c_b\rangle_a\right]\Phi + ||x_2 - c_b||_b^2 - ||x_2 - c_b||_a^2,$$

with c_a and c_b the centroids of $V(a)$ and $V(b)$, and associated inner products $\langle\cdot\rangle_{a,b}$. If this inequality can become true, then we are tunnelling. In this case, we set the local distance to ∞.

Another tunneling effect occurs at the higher level. In Figure 2, we still tunnel because the local metric at the tip of the bump yields a small Dirichlet region. We define *adjacency* of centroids if they are locally close. This is in general difficult to discover: it is the weakest point of the isomap algorithm [7]. Luckily, in our case, it is possible to define adjacency by watching the time curve $x(t)$: two Dirichlet regions a and b are adjacent if $\exists\tau$ such that $x(\tau)$ is in b and $x(\tau - 1)$ is in a or vice-versa. The local distance between two points in non-adjacent Dirichlet regions is ∞.

3.4. Geolen integration over a discrete manifold

We have now computed all local metrics by carefully avoiding the tuneling effect. It can be thought of as computing the $\mathrm{d}s^2$ lengths. To integrate the length as in (eq. 10), we need to compute the minimal integrals over a discrete manifold. The sampled manifold, endowed with local lengths, is an undirected weighted graph. From this graph, we would like to construct a fully connected graph with all minimal pairwise distances. This is done conveniently via an adaptation to undirected graphs of the Floyd-Warshall algorithm [8], which solves a problem called the All Shortest Paths problem (ASP).

3.5. Multi-dimensional scaling

Multidimensional scaling involves computing the SVD of a matrix with many zero eigenvalues. When the size of the matrix is greater than 10×10, this poses extraordinary numerical difficulties to standard linear algebra software. We add white noise to the observations, or a multiple of the identity to the B^* matrix:

$$\tilde{B}^* = B^* + \varepsilon I, \tag{29}$$

where $0 < \varepsilon \ll 1$ ensures strict positivity. This stabilizes the process.

3.6. Summary

We can therefore summarize the algorithm in three simple steps: computation of the sweeping metric at all quantization centroids, computation of the geolen reference, and transformation of the geolen into an isometric reference. The computation of the sweeping metric utilizes vector quantization: it will compute g^{kj} and g_{kj} for all Voronoi regions around each centroid. The geolen reference ultimately yields the Δ matrix, which relates each point to all other points. It is computed via Lagrangian optimization of small distances on nearby points. It is converted into global geolen via Floyd-Warshall. Finally, a spectral decomposition of the distance matrix yields the final coordinates $s(t)$.

4. EXPERIMENTS

The NIST 2002 Rich Transcription BN evaluation test set (RT-02) was selected for validation. It consists of six 10-minutes excerpts of Broadcast News. It was clustered using full, single Gaussian, BIC-penalized models [9]. MFCC coefficients were generated (13, excluding c0 and including energy), at a frame rate of 100Hz, and normalized with a centered sliding window cepstral mean normalization. Then, they were normalized using our novel algorithm. Results with MFCC parameters, SWAMP parameters, and MFCC parameters concatenated with SWAMP, are show on Table 1. Although our new parameters seem to improve clustering, it appears that they do not contain enough information in themselves to perform accurate clustering. We used NIST's RT-03S development scoring script SpkrEval-v20.pl. Thresholds and dimensions were roughly optimized. The quantizer used 12 clusters. To limit computational resources, the SWAMP frame rate was reduced to 10 Hz.

Features	Dimension	Error rate
MFCC	13	18.58%
SWAMP	13	38.61%
MFCC+SWAMP	18	17.52%

Table 1. NIST Speaker Error with different frontends

5. CONCLUSION AND FURTHER WORK

In this paper, we define a sound theoretical framework for natural isometric frontends based on differential geometry. It combines features of Levin [2] and Isomap [7]; also, it adds many key elements including sufficient conditions for noise robustness, tunnelling prevention, naturalness, and ergodicity. The resulting parameterization is invariant under wide-sense stationary transformations and quasi-stationary noise.

We have used the Riemannian sweeping metric in this paper. It is a convenient choice. However, frontends typically use a non-Riemannian dualistic structure (time, log-spectrum, and cepstrum). Therefore, further work will concentrate on non-Riemannian dualistic structures based on information geometric inference [1].

6. REFERENCES

[1] S. Amari and H. Nagaoka, *Methods of Information Geometry*, vol. 191 of *Translations of Mathematical Monographs*, AMS / Oxford UP, 2000.

[2] D. N. Levin, "Blind Normalization of Speech from Different Channels and Speakers," in *Proc. of ICSLP*, Sep. 2002, pp. 1425–1428.

[3] E. Schrödinger, *Space Time Structure*, Cambridge UP, 1963.

[4] F. W. Young and R. M. Hamer, *Multidimensional Scaling: History, Theory and Applications*, Erlbaum, N. Y., 1987.

[5] M. Sharir and A. Schorr, "On shortest paths in polyhedral spaces," *SIAM J. Comput.*, vol. 15, pp. 193–215, 1986.

[6] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge UP, 1992.

[7] J. B. Tennenbaum, V. de Silva, and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol. 290, no. 5500, Dec. 2000.

[8] T. H. Cormen (Ed.), C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, 2nd edition, 2001.

[9] Y. Moh, P. Nguyen, and J.-C. Junqua, "Towards Domain Independent Speaker Clustering," in *Proc. of ICASSP*, Apr 2003, p. To appear.